Paper 01-2017

# Rainfall in India: A Time Series Forecasting

Nevil Bruno, Pankhuri Dwivedi, Priyanka Sharma, Singapore Management University


## ABSTRACT

Agriculture plays an important role in the Indian economy and has a considerable share in the GDP. Of all the factors that affect the agricultural output in the country, Rainfall is one of the major climatic parameters and is also a major influencing factor for crop production. Crop agriculture practices of a region are normally dependent on the precipitation pattern of that area, especially the 'kharif' (monsoon) crops. The aim of the present study is to analyze rainfall time series over a wide time interval and a wide area, detecting potential trends. To achieve this goal, we have used seasonal rainfall data for a period of 15 years and used the same to gain insights on the rainfall patterns across different regions in the country. The dataset used for the project is extracted from the Open Government Data (OGD) Platform India. The model aims to predict the monthly rainfall for the country using time series analysis techniques. For exploring the dataset, we've used SAS JMP Pro and to perform the time series analysis, we've used SAS E-Miner tools. We have compared the different model's performance based on the R2, AIC & SBC values. Based on our analysis, the SARIMA model configuration (1,0,1) (1,1,1)12 was chosen as the final prediction model.


## INTRODUCTION

The agricultural practices and crop yields of India are heavily dependent on the climatic factors like rainfall. India ranks first among the rainfed agricultural countries of the world in terms of both extent and value of produce however, unlike irrigated agriculture, rain fed farming is usually diverse and risk prone. The monsoon season is the principal rain bearing season and a small variation in the timing and the quantity of monsoon rainfall has the potential to adversely impact the agricultural outcome. A prior knowledge of monsoon behavior can help Indian farmers and policy makers to take advantage of rain water and to minimize crop damage and human hardship during adverse monsoons.

## OBJECTIVE

This paper aims to predict monthly rainfall for India over a period of 5 years (2015-2019) obtained through time-series analysis techniques by utilizing 15 years (2000-2014) of rainfall data, extracted from Open Government Data (OGD) Platform India (data.gov.in).

## DATA PREPARATION

The original dataset consisted of monthly, annual and quarterly rainfall of the 36 subdivisions in India for a period of 115 years, 1901 – 2015. This data was subset to extract the monthly rainfall of the 36 subdivisions for 15 years from 2000-2014.

| | SUBDIVISION | JAN | FEB | MAR | APR | MAY | JUN | JUL | AUG | SEP | OCT | NOV | DEC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ANDAMAN & NICOBAR ISLANDS | 53 | 59 | 171.3 | 218.1 | 422.8 | 357 | 176.3 | 460.8 | 250.1 | 321.2 | 158.3 | 115.2 |
| 2 | ANDAMAN & NICOBAR ISLANDS | 89 | 15.7 | 143.3 | 30.1 | 705.3 | 370.7 | 341.3 | 469 | 334.4 | 267.6 | 222.6 | 91.8 |
| 3 | ANDAMAN & NICOBAR ISLANDS | 10.6 | 0 | 11.5 | 100.2 | 366.7 | 358.3 | 317.4 | 429.8 | 420 | 169 | 306.7 | 129.9 |
| 4 | ANDAMAN & NICOBAR ISLANDS | 44.3 | 7.9 | 149.2 | 19.4 | 296.3 | 159.9 | 494.9 | 379.4 | 371.9 | 310.4 | 74.1 | 48 |
| 5 | ANDAMAN & NICOBAR ISLANDS | 54.5 | 35.9 | 36.5 | 41.6 | 505.1 | 423.9 | 378.9 | 308.7 | 280.7 | 223.9 | 169.9 | 0.4 |
| 6 | ANDAMAN & NICOBAR ISLANDS | 0 | 0 | 20.3 | 51.1 | 305.3 | 452.4 | 429.3 | 311.1 | 507.5 | 293.5 | 300.1 | 283.9 |
| 7 | ANDAMAN & NICOBAR ISLANDS | 16.3 | 14.4 | 48.9 | 163.7 | 321.4 | 366 | 182.7 | 219.6 | 546.5 | 374.7 | 76.1 | 74.4 |
| 8 | ANDAMAN & NICOBAR ISLANDS | 4.9 | 1.6 | 5 | 54.6 | 370 | 378.3 | 463.2 | 465.3 | 486.4 | 209.4 | 223.9 | 85.6 |
| 9 | ANDAMAN & NICOBAR ISLANDS | 9.9 | 67.7 | 115.8 | 216.1 | 545.5 | 457.8 | 511.2 | 482.1 | 332 | 243.7 | 321.1 | 72 |
| 10 | ANDAMAN & NICOBAR ISLANDS | 24.5 | 6.3 | 44.2 | 136.5 | 313.1 | 633.5 | 297.3 | 351.5 | 344.7 | 272.8 | 66.2 | 48 |
| 11 | ANDAMAN & NICOBAR ISLANDS | 101.7 | 8 | 0.7 | 12.5 | 319 | 448.9 | 521.9 | 563.8 | 263.3 | 402.4 | 268.5 | 246.4 |
| 12 | ANDAMAN & NICOBAR ISLANDS | 265.9 | 84.8 | 272.8 | 111.4 | 326.5 | 383.2 | 583.2 | 441.5 | 757.1 | 212.3 | 150.8 | 238.5 |
| 13 | ANDAMAN & NICOBAR ISLANDS | 119.9 | 45.6 | 30.9 | 55.8 | 533.9 | 458.2 | 317.3 | 369.6 | 868.9 | 209.7 | 300.5 | 187.3 |
| 14 | ANDAMAN & NICOBAR ISLANDS | 67.1 | 37.6 | 43 | 46.3 | 509.3 | 777 | 564.8 | 336.7 | 473.6 | 455.8 | 354.2 | 92.3 |
| 15 | ANDAMAN & NICOBAR ISLANDS | 41.9 | 8.6 | 0 | 11.1 | 238 | 416.6 | 467.6 | 321.6 | 412.9 | 402.6 | 201.2 | 100.4 |
| 16 | ARUNACHAL PRADESH | 54.1 | 47.1 | 139.9 | 293.8 | 267.2 | 459.8 | 395.4 | 387.4 | 407.4 | 81.3 | 53.9 | 9.6 |
| 17 | ARUNACHAL PRADESH | 53.1 | 66.6 | 134.9 | 229.9 | 195.6 | 277.2 | 302.6 | 279.9 | 288.1 | 173.8 | 19.3 | 14.9 |
| 18 | ARUNACHAL PRADESH | 74.2 | 37.2 | 126.3 | 248.2 | 197.1 | 396.8 | 604 | 290.7 | 257.9 | 94 | 45 | 17.9 |
| 19 | ARUNACHAL PRADESH | 22.9 | 83.2 | 109.6 | 182 | 173.5 | 424.3 | 645.3 | 327.2 | 259.9 | 182.6 | 25 | 16.2 |
| 20 | ARUNACHAL PRADESH | 38 | 39.1 | 175.5 | 210.2 | 298.7 | 402.9 | 654.3 | 243 | 278.5 | 184.8 | 5.6 | 15.2 |
| 21 | ARUNACHAL PRADESH | 48.4 | 167.6 | 229.5 | 195.3 | 179.8 | 269.3 | 430.8 | 400 | 243.6 | 139.3 | 28.6 | 3.3 |
| 22 | ARUNACHAL PRADESH | 6 | 103.7 | 63.3 | 202.7 | 321.7 | 520.4 | 382.2 | 227.6 | 263.2 | 77.2 | 69.7 | 21.7 |
| 23 | ARUNACHAL PRADESH | 13.4 | 97.4 | 48.1 | 292.4 | 250.4 | 530.2 | 761 | 364.6 | 529.3 | 102.6 | 24.3 | 6.9 |
| 24 | ARUNACHAL PRADESH | 76.7 | 39.7 | 122.6 | 192.4 | 185 | 423.6 | 456.1 | 439.3 | 189.7 | 115.1 | 1.7 | 2.6 |
| 25 | ARUNACHAL PRADESH | 18 | 92.8 | 72.1 | 132.7 | 189.9 | 259.1 | 329.9 | 370.3 | 152.5 | 82.9 | 33.9 | 15.9 |
| 26 | ARUNACHAL PRADESH | 0.6 | 13.2 | 237.8 | 466.9 | 312.7 | 509.9 | 378 | 321.5 | 444.2 | 97.7 | 58.9 | 14.2 |
| 27 | ARUNACHAL PRADESH | 40 | 51.3 | 174.5 | 240.8 | 219.6 | 288.4 | 531.4 | 277.6 | 286.7 | 51.9 | 16.2 | 15.2 |
| 28 | ARUNACHAL PRADESH | 57.8 | 35.8 | 134.2 | 403.4 | 187.4 | 645.8 | 638.9 | 316 | 724.9 | 248.1 | 22 | 26.2 |
| 29 | ARUNACHAL PRADESH | 18.5 | 40.5 | 115.1 | 175.1 | 335.8 | 290 | 329.6 | 230.2 | 316.1 | 164.1 | 13.3 | 14.6 |
| 30 | ARUNACHAL PRADESH | 19 | 101.9 | 80.3 | 86.7 | 299 | 415.8 | 392.4 | 599.6 | 343 | 35.1 | 20.1 | 10.2 |
| 31 | ASSAM & MEGHALAYA | 20.5 | 13.7 | 65.9 | 238.5 | 375.1 | 540.2 | 313.3 | 513.2 | 299.9 | 83.9 | 22.2 | 0.4 |

**Figure 1. Monthly Rainfall Dataset**

To prepare the data for time series analysis, the table was transposed to obtain the monthly rainfall as a time series pattern for each subdivision, with 180 rows indicating the time-series and the subdivisions as columns.

| | Numerical Series | Series | YEAR | Label | ANDAMAN & NICOBAR ... | ARUNACHAL PRADESH | ASSAM & MEGHALAYA | NAGA MANI MIZO TRIPURA | SUB HIMALAYAN WEST BENGAL ... | GANGETIC WEST BENGAL | ORISSA | JH |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | JAN-00 | 2000 | JAN | 53 | 54.1 | 20.5 | 22.5 | 10.7 | 9.5 | 0.9 | |
| 2 | 2 | FEB-00 | 2000 | FEB | 59 | 47.1 | 13.7 | 26.1 | 36.2 | 57.9 | 33.3 | |
| 3 | 3 | MAR-00 | 2000 | MAR | 171.3 | 139.9 | 65.9 | 129.1 | 55.1 | 3.8 | 1.1 | |
| 4 | 4 | APR-00 | 2000 | APR | 218.1 | 293.8 | 238.5 | 214.2 | 185.1 | 54 | 21.3 | |
| 5 | 5 | MAY-00 | 2000 | MAY | 422.8 | 267.2 | 375.1 | 403.6 | 326.5 | 215 | 74.7 | |
| 6 | 6 | JUN-00 | 2000 | JUN | 357 | 459.8 | 540.2 | 325.4 | 649.6 | 205.6 | 233.7 | |
| 7 | 7 | JUL-00 | 2000 | JUL | 176.3 | 395.4 | 313.3 | 269.2 | 574.3 | 341.5 | 272.3 | |
| 8 | 8 | AUG-00 | 2000 | AUG | 460.8 | 387.4 | 513.2 | 465 | 498.1 | 178.4 | 273.6 | |
| 9 | 9 | SEP-00 | 2000 | SEP | 250.1 | 407.4 | 299.9 | 298.3 | 465.8 | 428.4 | 158 | |
| 10 | 10 | OCT-00 | 2000 | OCT | 321.2 | 81.3 | 83.9 | 185.2 | 82.9 | 76.2 | 26.6 | |
| 11 | 11 | NOV-00 | 2000 | NOV | 158.3 | 53.9 | 22.2 | 24.2 | 22.6 | 1.6 | 1.6 | |
| 12 | 12 | DEC-00 | 2000 | DEC | 115.2 | 9.6 | 0.4 | 1.1 | 1.2 | 0.1 | 0.1 | |
| 13 | 13 | JAN-01 | 2001 | JAN | 89 | 53.1 | 5.5 | 1.5 | 4 | 0.3 | 0.4 | |
| 14 | 14 | FEB-01 | 2001 | FEB | 15.7 | 66.6 | 39.7 | 46.9 | 20.5 | 1.9 | 1.4 | |
| 15 | 15 | MAR-01 | 2001 | MAR | 143.3 | 134.9 | 30.9 | 38.8 | 50.2 | 32.8 | 32.3 | |
| 16 | 16 | APR-01 | 2001 | APR | 30.1 | 229.9 | 198.6 | 83.6 | 134.8 | 49.4 | 28.7 | |
| 17 | 17 | MAY-01 | 2001 | MAY | 705.3 | 195.6 | 266.4 | 313.8 | 347.2 | 175.8 | 71 | |
| 18 | 18 | JUN-01 | 2001 | JUN | 370.7 | 277.2 | 399.8 | 478 | 472.2 | 384.2 | 336.1 | |
| 19 | 19 | JUL-01 | 2001 | JUL | 341.3 | 302.6 | 451.1 | 322.6 | 399 | 269.3 | 584.2 | |
| 20 | 20 | AUG-01 | 2001 | AUG | 469 | 279.9 | 295.6 | 263 | 424.5 | 248.7 | 380.9 | |
| 21 | 21 | SEP-01 | 2001 | SEP | 334.4 | 288.1 | 266.8 | 269 | 434.9 | 205.9 | 132 | |
| 22 | 22 | OCT-01 | 2001 | OCT | 267.6 | 173.8 | 201.4 | 266.4 | 282.8 | 163.6 | 86.8 | |
| 23 | 23 | NOV-01 | 2001 | NOV | 222.6 | 19.3 | 20.5 | 65.4 | 36.6 | 8.8 | 18.7 | |
| 24 | 24 | DEC-01 | 2001 | DEC | 91.8 | 14.9 | 2.3 | 0.2 | 5.9 | 0 | 0 | |
| 25 | 25 | JAN-02 | 2002 | JAN | 10.6 | 74.2 | 20.2 | 18.6 | 30.1 | 26.3 | 11.7 | |
| 26 | 26 | FEB-02 | 2002 | FEB | 0 | 37.2 | 5.7 | 3.9 | 10 | 1.5 | 0.2 | |
| 27 | 27 | MAR-02 | 2002 | MAR | 11.5 | 126.3 | 83.3 | 62.7 | 95.6 | 26.6 | 13.1 | |
| 28 | 28 | APR-02 | 2002 | APR | 100.2 | 248.2 | 279.9 | 143.8 | 237.1 | 77.4 | 25.1 | |
| 29 | 29 | MAY-02 | 2002 | MAY | 366.7 | 197.1 | 286.9 | 438.7 | 181.7 | 114.1 | 67.5 | |
| 30 | 30 | JUN-02 | 2002 | JUN | 358.3 | 396.8 | 502.5 | 351.6 | 407.2 | 347.7 | 182.4 | |

**Figure 2. Monthly Rainfall Time Series Dataset**

For our graphical visualization, we required a map of India with the region boundaries on JMP Pro. The shape file for India with the region boundaries was not available on any online library or forums. To make this shape file, JMP's 'Custom Map Creator' Add-In was installed and used. To plot out each boundary, a rainfall region map obtained from the Indian Meteorological Department's website was used as an underlying reference to plot the X and Y coordinated and mapping each shape to a region. The generated shape files were then placed in the maps folder in the JMP installer file. By doing so, the map shape file for the Indian Rainfall Region will be generated each time any graph plots involving the 36 regions are involved.

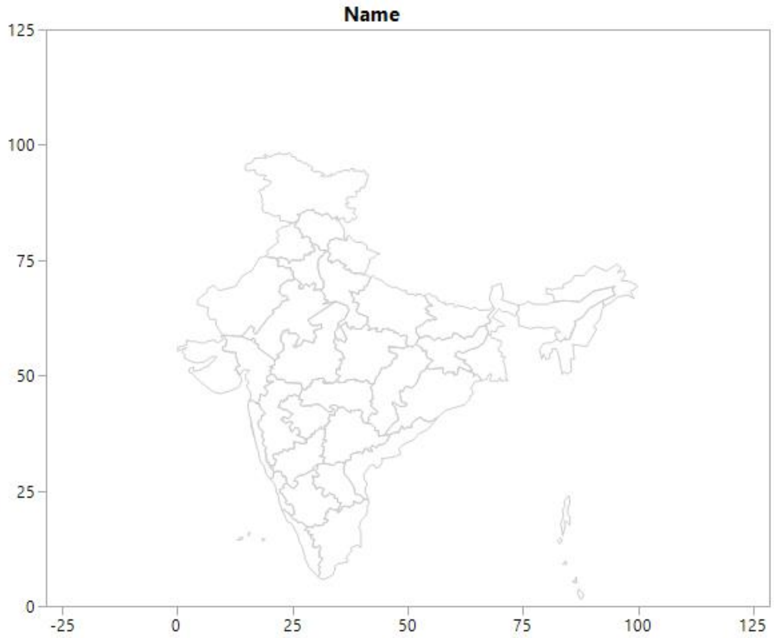



**Figure 3. Shape File of India Consisting of the 36 Rainfall Sub-Divisions**

## METHODOLOGY

### TIME SERIES CLUSTERING

#### Introduction to Time Series Clustering:

Time series clustering is to partition time series data into groups based on similarity or distance, so that time series in the same cluster are similar. Clustering time-series data has been used in diverse scientific areas to discover patterns which empower data analysts to extract valuable information from complex and massive datasets. In case of huge datasets, using supervised classification solutions is almost impossible, while clustering can solve this problem using un-supervised approaches. Hierarchical clustering, unlike k-means, is a deterministic algorithm, so we can't reuse the experimental methodology from the previous section exactly, however, we can do something very similar.

#### MODEL BUILDING

In this case, we have 36 regions for 180 weeks. Creating 36 predictive forecast models for 36 regions is redundant and time consuming. To ease the process of modelling, we employ clustering techniques to identify regions that share similar characteristics with respect to the amount of rainfall it receives in a period of 12 months. From here, we build forecasting models for the resulting clusters.

For our clustering, we will use SAS Enterprise Miner. We will be using a data set containing monthly rainfall data for the 36 regions for a period of 15 years (2000 - 2014). The data file is in a .csv format, so we will be using the File Import node. We import the .csv file into SAS EM.

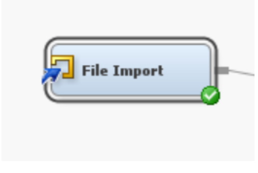
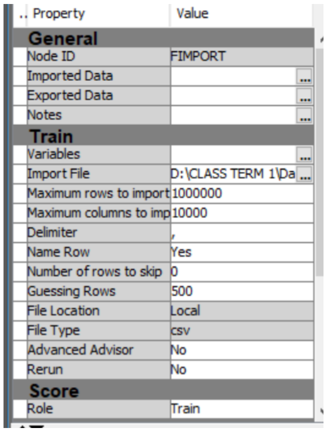



**Figure 4. SAS EM File Import Node**



| Property | Value |
|---|---|
| **General** | |
| Node ID | FIMPORT |
| Imported Data | ... |
| Exported Data | ... |
| Notes | ... |
| **Train** | |
| Variables | ... |
| Import File | D:\CLASS TERM 1\Da ... |
| Maximum rows to import | 1000000 |
| Maximum columns to imp | 10000 |
| Delimiter | , |
| Name Row | Yes |
| Number of rows to skip | 0 |
| Guessing Rows | 500 |
| File Location | Local |
| File Type | csv |
| Advanced Advisor | No |
| Rerun | No |
| **Score** | |
| Role | Train |

**Figure 5. SAS EM File Import Options Panel**

Now that we have the file imported on SAS EM, we can explore the variables using the metadata node. Using this node, we can explore and define the variable roles.

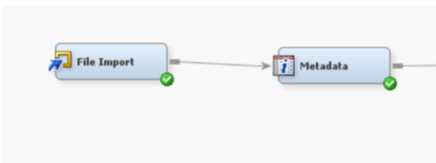
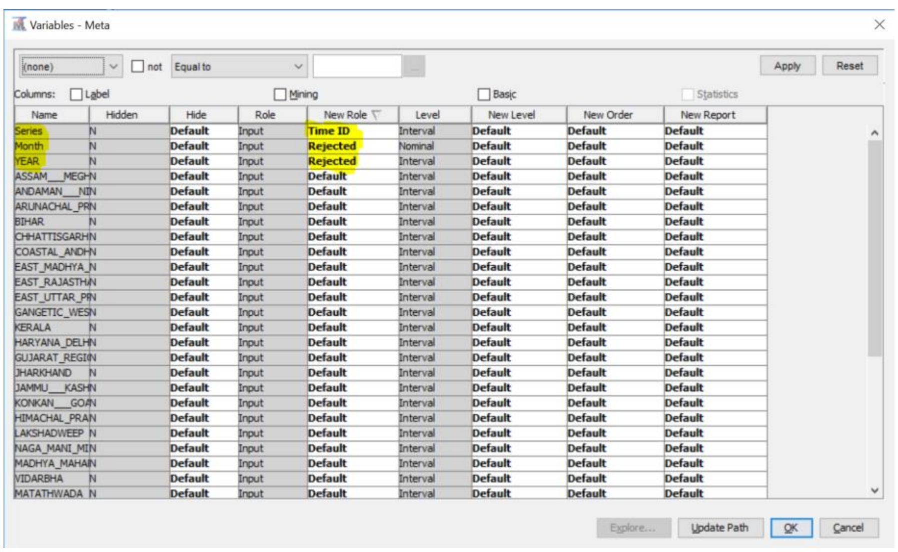



**Figure 6. SAS EM Metadata Node**

We Keep all the region columns as input. We exclude the Month and Year columns from our clustering by setting them to 'Rejected'. We set the series column to 'Time ID' role. We run this node to apply the new changes.




| Name | Hidden | Hide | Role | New Role | Level | New Level | New Order | New Report |
|---|---|---|---|---|---|---|---|---|
| Series | N | Default | Input | Time ID | Interval | Default | Default | Default |
| Month | N | Default | Input | Rejected | Nominal | Default | Default | Default |
| YEAR | N | Default | Input | Rejected | Interval | Default | Default | Default |
| ASSAM__MEGH | N | Default | Input | Default | Interval | Default | Default | Default |
| ANDAMAN__NI | N | Default | Input | Default | Interval | Default | Default | Default |
| ARUNACHAL_PR | N | Default | Input | Default | Interval | Default | Default | Default |
| BIHAR | N | Default | Input | Default | Interval | Default | Default | Default |
| CHHATTISGARH | N | Default | Input | Default | Interval | Default | Default | Default |
| COASTAL_ANDH | N | Default | Input | Default | Interval | Default | Default | Default |
| EAST_MADHYA_ | N | Default | Input | Default | Interval | Default | Default | Default |
| EAST_RAJASTH | N | Default | Input | Default | Interval | Default | Default | Default |
| EAST_UTTAR_P | N | Default | Input | Default | Interval | Default | Default | Default |
| GANGETIC_WES | N | Default | Input | Default | Interval | Default | Default | Default |
| KERALA | N | Default | Input | Default | Interval | Default | Default | Default |
| HARYANA_DELH | N | Default | Input | Default | Interval | Default | Default | Default |
| GUJARAT_REGI | N | Default | Input | Default | Interval | Default | Default | Default |
| JHARKHAND | N | Default | Input | Default | Interval | Default | Default | Default |
| JAMMU__KASH | N | Default | Input | Default | Interval | Default | Default | Default |
| KONKAN__GOA | N | Default | Input | Default | Interval | Default | Default | Default |
| HIMACHAL_PRA | N | Default | Input | Default | Interval | Default | Default | Default |
| LAKSHADWEEP | N | Default | Input | Default | Interval | Default | Default | Default |
| NAGA_MANI_MI | N | Default | Input | Default | Interval | Default | Default | Default |
| MADHYA_MAHA | N | Default | Input | Default | Interval | Default | Default | Default |
| VIDARBHA | N | Default | Input | Default | Interval | Default | Default | Default |
| MATATHWADA | N | Default | Input | Default | Interval | Default | Default | Default |

**Figure 7. SAS EM Metadata Variables Selection**

Next, we use the TS Data Preparation Node to set the differencing settings. This is done since we have seasonal time series data. We run the node after we set the differencing values.

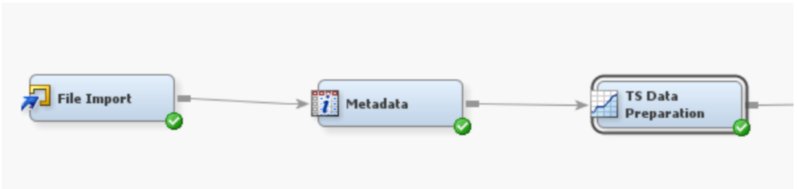
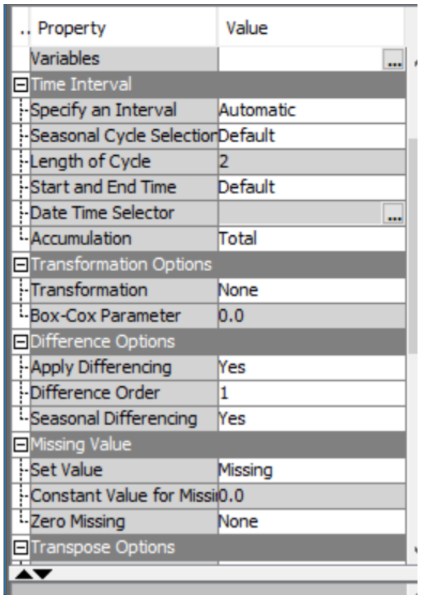



**Figure 8. SAS EM TS Data Preparation Node**



| Property | Value |
|---|---|
| Variables | ... |
| Time Interval | |
| Specify an Interval | Automatic |
| Seasonal Cycle Selection | Default |
| Length of Cycle | 2 |
| Start and End Time | Default |
| Date Time Selector | ... |
| Accumulation | Total |
| Transformation Options | |
| Transformation | None |
| Box-Cox Parameter | 0.0 |
| Difference Options | |
| Apply Differencing | Yes |
| Difference Order | 1 |
| Seasonal Differencing | Yes |
| Missing Value | |
| Set Value | Missing |
| Constant Value for Missi | 0.0 |
| Zero Missing | None |
| Transpose Options | |

**Figure 9. SAS EM TS Data Preparation Options Panel**

Now we use the TS Similarity node. This node is responsible for the clustering.

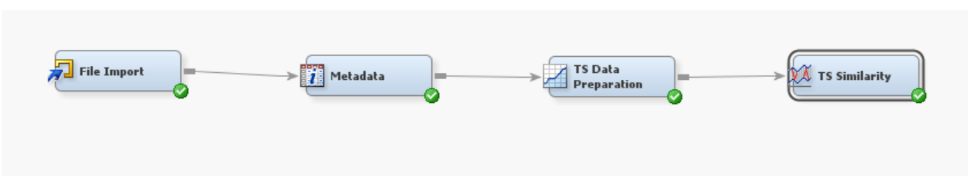



**Figure 10. SAS EM TS Similarity Node**

Since we are dealing with Time series clustering, we set the Hierarchical clustering option to default. We also have the option to set the number of clusters. The minimum number of clusters for any clustering process is 3. On setting the number of clusters to 3, we obtained three clusters that were not significantly different and separated. On selecting the number of clusters as 5, the 5th cluster had only 2 regions. Due to these problems, we rejected these clusters.

On setting the number of clusters to 4, we can obtain 4 well defined and distinct clusters.

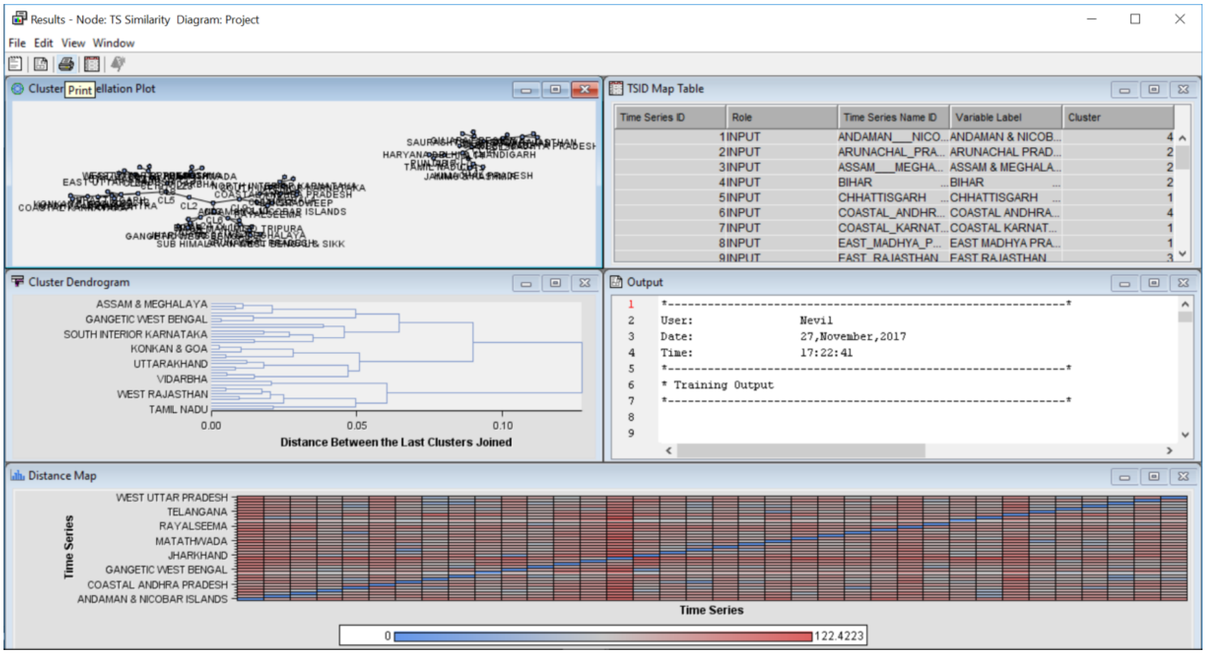


Figure 11. SAS EM TS Similarity Results

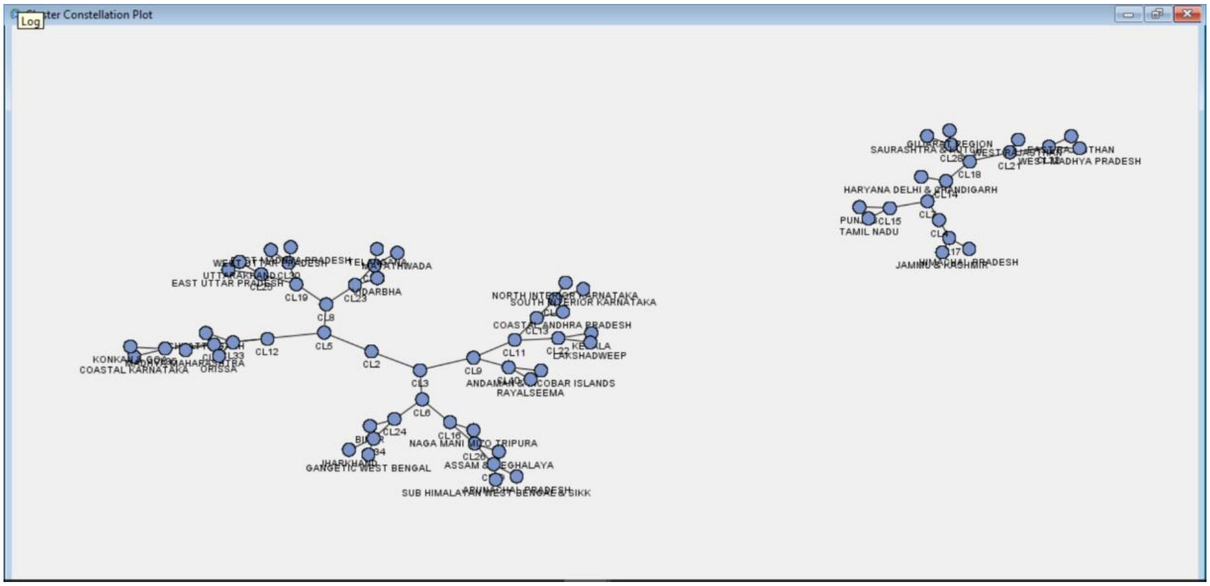


Figure 12. SAS EM Cluster Constellation Plot

### Cluster Details

| TIME SERIES ID | ROLE | TIME SERIES NAME ID | VARIABLE NAME | CLUSTER |
|---|---|---|---|---|
| 1 | INPUT | ANDAMAN___NICOBAR_ISLANDS | ANDAMAN & NICOBAR ISLANDS | 4 |
| 2 | INPUT | ARUNACHAL_PRADESH | ARUNACHAL PRADESH | 2 |
| 3 | INPUT | ASSAM___MEGHALAYA | ASSAM & MEGHALAYA | 2 |
| 4 | INPUT | BIHAR | BIHAR | 2 |

| | | | | |
|---|---|---|---|---|
| 5 | INPUT | CHHATTISGARH | CHHATTISGARH | 1 |
| 6 | INPUT | COASTAL_ANDHRA_PRADESH | COASTAL ANDHRA PRADESH | 4 |
| 7 | INPUT | COASTAL_KARNATAKA | COASTAL KARNATAKA | 1 |
| 8 | INPUT | EAST_MADHYA_PRADESH | EAST MADHYA PRADESH | 1 |
| 9 | INPUT | EAST_RAJASTHAN | EAST RAJASTHAN | 3 |
| 10 | INPUT | EAST_UTTAR_PRADESH | EAST UTTAR PRADESH | 1 |
| 11 | INPUT | GANGETIC_WEST_BENGAL | GANGETIC WEST BENGAL | 2 |
| 12 | INPUT | GUJARAT_REGION | GUJARAT REGION | 3 |
| 13 | INPUT | HARYANA_DELHI___CHANDIGARH | HARYANA DELHI & CHANDIGARH | 3 |
| 14 | INPUT | HIMACHAL_PRADESH | HIMACHAL PRADESH | 3 |
| 15 | INPUT | JAMMU___KASHMIR | JAMMU & KASHMIR | 3 |
| 16 | INPUT | JHARKHAND | JHARKHAND | 2 |
| 17 | INPUT | KERALA | KERALA | 4 |
| 18 | INPUT | KONKAN___GOA | KONKAN & GOA | 1 |
| 19 | INPUT | LAKSHADWEEP | LAKSHADWEEP | 4 |
| 20 | INPUT | MADHYA_MAHARASHTRA | MADHYA MAHARASHTRA | 1 |
| 21 | INPUT | MATATHWADA | MATATHWADA | 1 |
| 22 | INPUT | NAGA_MANI_MIZO_TRIPURA | NAGA MANI MIZO TRIPURA | 2 |
| 23 | INPUT | NORTH_INTERIOR_KARNATAKA | NORTH INTERIOR KARNATAKA | 4 |
| 24 | INPUT | ORISSA | ORISSA | 1 |
| 25 | INPUT | PUNJAB | PUNJAB | 3 |
| 26 | INPUT | RAYALSEEMA | RAYALSEEMA | 4 |
| 27 | INPUT | SAURASHTRA___KUTCH | SAURASHTRA & KUTCH | 3 |
| 28 | INPUT | SOUTH_INTERIOR_KARNATAKA | SOUTH INTERIOR KARNATAKA | 4 |
| 29 | INPUT | SUB_HIMALAYAN_WEST_BENGAL___SIKK | SUB HIMALAYAN WEST BENGAL & SIKK | 2 |
| 30 | INPUT | TAMIL_NADU | TAMIL NADU | 3 |
| 31 | INPUT | TELANGANA | TELANGANA | 1 |
| 32 | INPUT | UTTARAKHAND | UTTARAKHAND | 1 |
| 33 | INPUT | VIDARBHA | VIDARBHA | 1 |
| 34 | INPUT | WEST_MADHYA_PRADESH | WEST MADHYA PRADESH | 3 |
| 35 | INPUT | WEST_RAJASTHAN | WEST RAJASTHAN | 3 |
| 36 | INPUT | WEST_UTTAR_PRADESH | WEST UTTAR PRADESH | 1 |

We use this cluster table and select the regions and group them accordingly in JMP Pro to further analyze the clusters.

We group the regions as per the clusters in JMP Pro and calculate the monthly average rainfall for the four clusters. Plotting the 4 cluster averages for a period of 12 months, we can analyze and find unique features of each cluster.

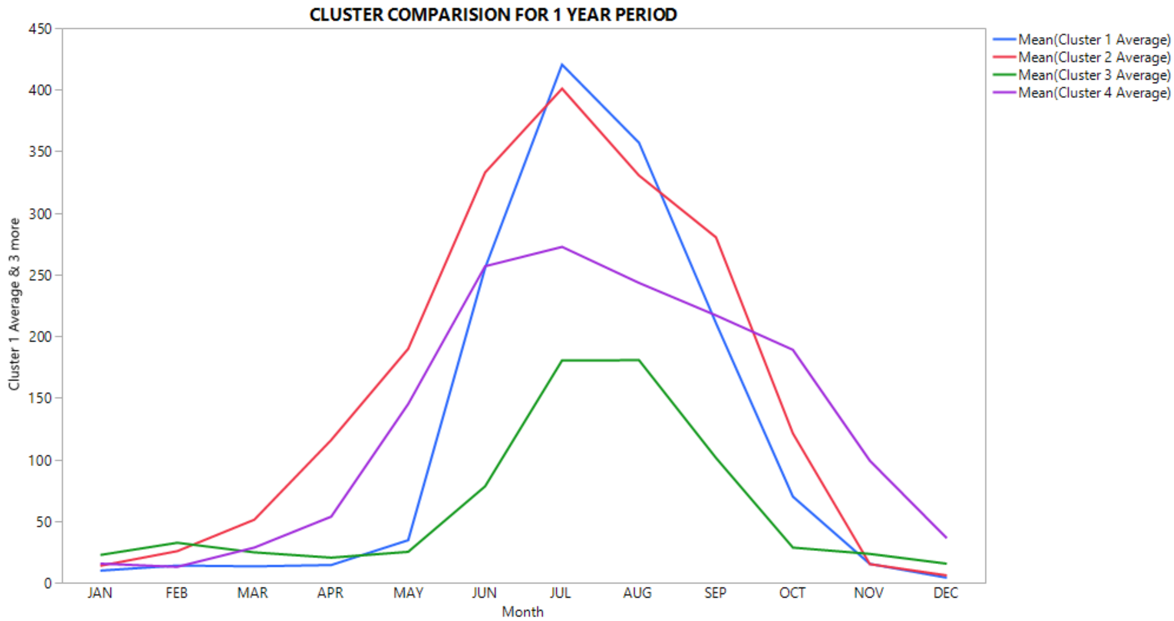



**Figure 13. Cluster Comparison**

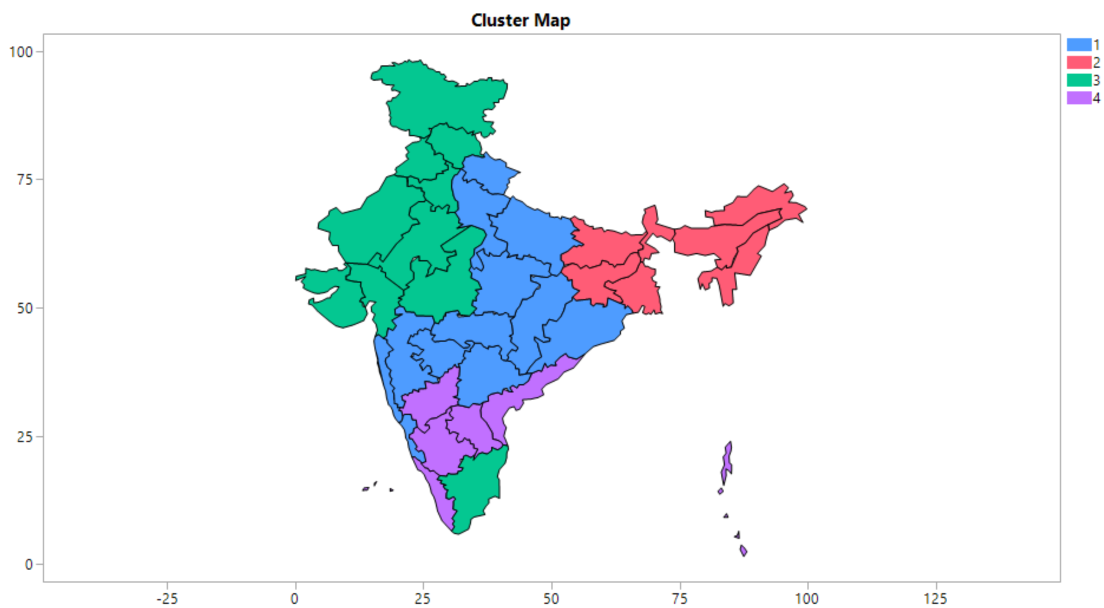



**Figure 14. Cluster Map for India**

From our analysis, we can identify the following key distinctive characteristics for each cluster:

**CLUSTER 1**

Covers 12 regions. Central and some northern parts of India

Receives moderate rainfall from June to September

Maximum rainfall in July

**CLUSTER 2**

Covers 7 regions. Eastern parts of India.

Receives high rainfall from May to October.

Moderate winter rains.

**CLUSTER 3**

Covers 10 Regions. North, north western regions, and Tamil Nadu

Dry throughout the year with few moderate showers in July and November

Receives least rainfall compared to the other three cluster regions

**CLUSTER 4**

Covers 7 regions. Mainly South India and the two island territories

Receives rainfall for more than half a year (May to January)

Highest average rainfall amongst the 4 clusters

Using the data from the 4 cluster averages, we proceed to make a Seasonal Arima forecasting model.

## TIME SERIES FORECASTING

**Introduction to Time Series Forecasting:**

A time series analysis often exhibits four main components such as trends, seasonality, cycles and irregular fluctuations. It is represented by the equation:

$\boldsymbol{Yt} = \boldsymbol{Tt} + \boldsymbol{St} + \boldsymbol{C}\mathrm{t} + \boldsymbol{I}\mathrm{t}$

where $Yt$ is the observed time series, $Tt$ is the trend component, $St$ is the seasonal component, $Ct$ is the cyclical component, and $It$ is the irregular component.

For our project we have used the Box-Jenkins methodology which applies ARMA, ARIMA or SARIMA to establish the best fit of a time series historical values to make forecasts. This paper describes the Box-Jenkins time series Seasonal ARIMA (Auto Regressive Integrated Moving Average) approach for prediction of rainfall on a monthly scale.

The methodology consists of four stages namely model identification, estimation of model parameters, diagnostic checking for the identified model appropriateness for modelling, and application of the model (i.e. forecasting).

In the Identification stage, tentative values of p, d, q and P, D, Q(Seasonal) were chosen. Coefficients of variables used in model were estimated.

For the estimation of the model, diagnostic checks were made to determine, whether the model selected adequately describes the given time series. Any inadequacies discovered might suggest an alternative form of the model, and whole iterative cycle of identification, estimation and application was repeated until a satisfactory model was obtained.

Once the appropriate model was determined, it was applied to the existing series to predict the rainfall for the next 60 periods, i.e. 5 years.

Since, an average rainfall sequence has 12 cycles of seasonal change on trend, the season series differencing was taken into consideration i.e. while choosing the model parameters, we accounted for the season differencing.

**MODEL BUILDING**

Since, the original sequence had one order difference for a period of 12 months of the seasonal difference, the value d=0, D=1, S=12 was ascertained by observing the ACF (Auto-Correlation Function) and PACF (Partial Auto-Correlation Function).

(See Figure 14 as an example for Cluster 1 Time Series)

**Time Series Basic Diagnostics**

| Lag | AutoCorr | -.8-.6-.4-.2 0 .2 .4 .6 .8 | Ljung-Box Q | p-Value | Lag | Partial | -.8-.6-.4-.2 0 .2 .4 .6 .8 |
|---|---|---|---|---|---|---|---|
| 0 | 1.0000 | | . | . | 0 | 1.0000 | |
| 1 | 0.6794 | | 84.4887 | <.0001* | 1 | 0.6794 | |
| 2 | 0.2174 | | 93.1899 | <.0001* | 2 | -0.4536 | |
| 3 | -0.2049 | | 100.960 | <.0001* | 3 | -0.2606 | |
| 4 | -0.4284 | | 135.128 | <.0001* | 4 | -0.0767 | |
| 5 | -0.4943 | | 180.871 | <.0001* | 5 | -0.2004 | |
| 6 | -0.4851 | | 225.182 | <.0001* | 6 | -0.2961 | |
| 7 | -0.4639 | | 265.931 | <.0001* | 7 | -0.3867 | |
| 8 | -0.3941 | | 295.506 | <.0001* | 8 | -0.4172 | |
| 9 | -0.1868 | | 302.187 | <.0001* | 9 | -0.3525 | |
| 10 | 0.2120 | | 310.847 | <.0001* | 10 | -0.0330 | |
| 11 | 0.6375 | | 389.617 | <.0001* | 11 | 0.1997 | |
| 12 | 0.8498 | | 530.438 | <.0001* | 12 | 0.2968 | |
| 13 | 0.6556 | | 614.761 | <.0001* | 13 | -0.0156 | |
| 14 | 0.2060 | | 623.138 | <.0001* | 14 | -0.1850 | |
| 15 | -0.1925 | | 630.493 | <.0001* | 15 | 0.0062 | |
| 16 | -0.4019 | | 662.757 | <.0001* | 16 | 0.0955 | |
| 17 | -0.4641 | | 706.050 | <.0001* | 17 | -0.0000 | |
| 18 | -0.4523 | | 747.425 | <.0001* | 18 | -0.0160 | |
| 19 | -0.4245 | | 784.091 | <.0001* | 19 | -0.0052 | |
| 20 | -0.3657 | | 811.470 | <.0001* | 20 | -0.0672 | |
| 21 | -0.1792 | | 818.085 | <.0001* | 21 | -0.0847 | |
| 22 | 0.1865 | | 825.299 | <.0001* | 22 | -0.0240 | |
| 23 | 0.5925 | | 898.542 | <.0001* | 23 | 0.0375 | |
| 24 | 0.7899 | | 1029.57 | <.0001* | 24 | 0.0622 | |
| 25 | 0.6054 | | 1107.04 | <.0001* | 25 | -0.1209 | |

**Figure 14. Cluster 1 Time Series**

To make an initial guess, we primarily determined q=1 or 2, p=1 or 2. To set orders of the model methods, the minimum AIC (Akaike info criterion) criterion, minimum SBC (Schwarz Bayesian criterion), and the adjusted R squared were used. For this purpose, we respectively verify models SARIMA (1,0,1) (1,1,1)12, SARIMA (1,0,2) (1,1,1)12, SARIMA (2,0,1) (1,1,1)12 by applying the criteria for optimizing the combination.

**Cluster 1 Model Comparison:**

**Model Comparison**

| Report | Graph | Model | DF | Variance | AIC | SBC | RSquare | -2LogLH | Weights | .2 .4 .6 .8 | MAPE | MAE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ☑ | ☐ | Seasonal ARIMA(1, 0, 1)(1, 1, 1)12 | 163 | 2137.7221 | 1787.2808 | 1802.9006 | 0.902 | 1777.2808 | 0.549371 | | 114.04461 | 28.829789 |
| ☑ | ☐ | Seasonal ARIMA(1, 0, 2)(1, 1, 1)12 | 162 | 2149.0706 | 1789.0342 | 1807.7780 | 0.902 | 1777.0342 | 0.228626 | | 114.65233 | 28.920675 |
| ☑ | ☐ | Seasonal ARIMA(2, 0, 1)(1, 1, 1)12 | 162 | 2149.325 | 1789.0930 | 1807.8368 | 0.902 | 1777.093 | 0.222004 | | 115.00251 | 28.921146 |
| ☑ | ☐ | Seasonal ARIMA(1, 0, 1)(1, 0, 1)12 | 175 | 2173.614 | 1947.4043 | 1963.3691 | 0.833 | 1937.4043 | 0.000000 | | 208.18810 | 35.697222 |
| ☑ | ☐ | Seasonal ARIMA(2, 0, 1)(1, 0, 1)12 | 174 | 2180.8853 | 1948.8520 | 1968.0098 | 0.834 | 1936.852 | 0.000000 | | 201.66394 | 35.558000 |

**Figure 15. Model Comparison**

Based on the AIC and SBC values, the appropriate form of the original time series, SARIMA (2,0,2) (1,1,1)12 was selected as the final prediction model.

Selected Model Parameter Estimates:

**Model: Seasonal ARIMA(1, 0, 1)(1, 1, 1)12**

**Model Summary**

| | | | |
|---|---|---|---|
| DF | 163 | Stable | Yes |
| Sum of Squared Errors | 348448.704 | Invertible | Yes |
| Variance Estimate | 2137.72211 | | |
| Standard Deviation | 46.235507 | | |
| Akaike's 'A' Information Criterion | 1787.28081 | | |
| Schwarz's Bayesian Criterion | 1802.90063 | | |
| RSquare | 0.90150418 | | |
| RSquare Adj | 0.8990871 | | |
| MAPE | 114.04461 | | |
| MAE | 28.8297892 | | |
| -2LogLikelihood | 1777.28081 | | |

**Parameter Estimates**

| Term | Factor | Lag | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|---|---|
| AR1,1 | 1 | 1 | -0.4508381 | 0.4195839 | -1.07 | 0.2842 |
| AR2,12 | 2 | 12 | -0.1989805 | 0.0908524 | -2.19 | 0.0299* |
| MA1,1 | 1 | 1 | -0.3157749 | 0.4451810 | -0.71 | 0.4791 |
| MA2,12 | 2 | 12 | 0.8193776 | 0.0810523 | 10.11 | <.0001* |
| Intercept | 1 | 0 | 0.7182986 | 0.7582560 | 0.95 | 0.3449 |

| Constant Estimate | Mu |
|---|---|
| 1.24949954 | 0.71829861 |

**Forecast**

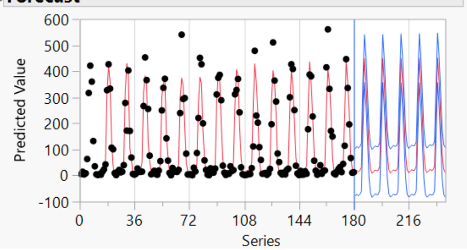



**Figure 16. Selected Model Parameter Estimates**

The model selected has an $R^2$ of 0.9 and an Adjusted $R^2$ of 0.89.

The greater value of the adjusted R-squared represents better model fitting.

For the verification of the model we perform the test for the white-noise, i.e. check the residuals to account for the stability of the model. If the residuals are not a white-noise sequence, it means that we must perform the modelling again to achieve a better model.

On observing the ACF and PACF of the model for Cluster 1 (Figure 17), we can conclude that the residuals are white noise and thus we can apply the selected model for forecasting.

| Lag | AutoCorr | -.8-.6-.4-.2 0 .2 .4 .6 .8 | Ljung-Box Q | p-Value | Lag | Partial | -.8-.6-.4-.2 0 .2 .4 .6 .8 |
|---|---|---|---|---|---|---|---|
| 0 | 1.0000 | | . | . | 0 | 1.0000 | |
| 1 | 0.0060 | | 0.0061 | 0.9377 | 1 | 0.0060 | |
| 2 | 0.0093 | | 0.0209 | 0.9896 | 2 | 0.0092 | |
| 3 | 0.0261 | | 0.1392 | 0.9868 | 3 | 0.0260 | |
| 4 | -0.0153 | | 0.1797 | 0.9962 | 4 | -0.0157 | |
| 5 | -0.0543 | | 0.6971 | 0.9831 | 5 | -0.0547 | |
| 6 | -0.0108 | | 0.7177 | 0.9941 | 6 | -0.0106 | |
| 7 | -0.0025 | | 0.7188 | 0.9982 | 7 | -0.0005 | |
| 8 | -0.0311 | | 0.8910 | 0.9988 | 8 | -0.0283 | |
| 9 | -0.0184 | | 0.9521 | 0.9995 | 9 | -0.0193 | |
| 10 | 0.1643 | | 5.8307 | 0.8293 | 10 | 0.1627 | |
| 11 | -0.0142 | | 5.8675 | 0.8821 | 11 | -0.0158 | |
| 12 | -0.0353 | | 6.0954 | 0.9112 | 12 | -0.0406 | |
| 13 | -0.0486 | | 6.5311 | 0.9247 | 13 | -0.0616 | |
| 14 | -0.1136 | | 8.9235 | 0.8359 | 14 | -0.1123 | |
| 15 | -0.1781 | | 14.8422 | 0.4628 | 15 | -0.1661 | |
| 16 | -0.0502 | | 15.3149 | 0.5017 | 16 | -0.0503 | |
| 17 | -0.0379 | | 15.5866 | 0.5533 | 17 | -0.0374 | |
| 18 | -0.0107 | | 15.6082 | 0.6199 | 18 | -0.0006 | |
| 19 | -0.0095 | | 15.6257 | 0.6821 | 19 | -0.0156 | |
| 20 | -0.0227 | | 15.7248 | 0.7335 | 20 | -0.0729 | |
| 21 | -0.0557 | | 16.3266 | 0.7510 | 21 | -0.0753 | |
| 22 | 0.0550 | | 16.9184 | 0.7678 | 22 | 0.0496 | |
| 23 | 0.0916 | | 18.5699 | 0.7259 | 23 | 0.1017 | |
| 24 | -0.0072 | | 18.5802 | 0.7740 | 24 | 0.0231 | |
| 25 | 0.0335 | | 18.8044 | 0.8064 | 25 | 0.0787 | |

**Figure 17. Selected Model Parameter Estimates**

## FORECAST RESULTS

The following graphs show the forecast results of Actual Vs. the Predicted Rainfall. The predicted results when compared to the actual for the same time period have a good overlap fit, with a few unexpected spikes which can be due to other factors affecting the rainfall patterns:

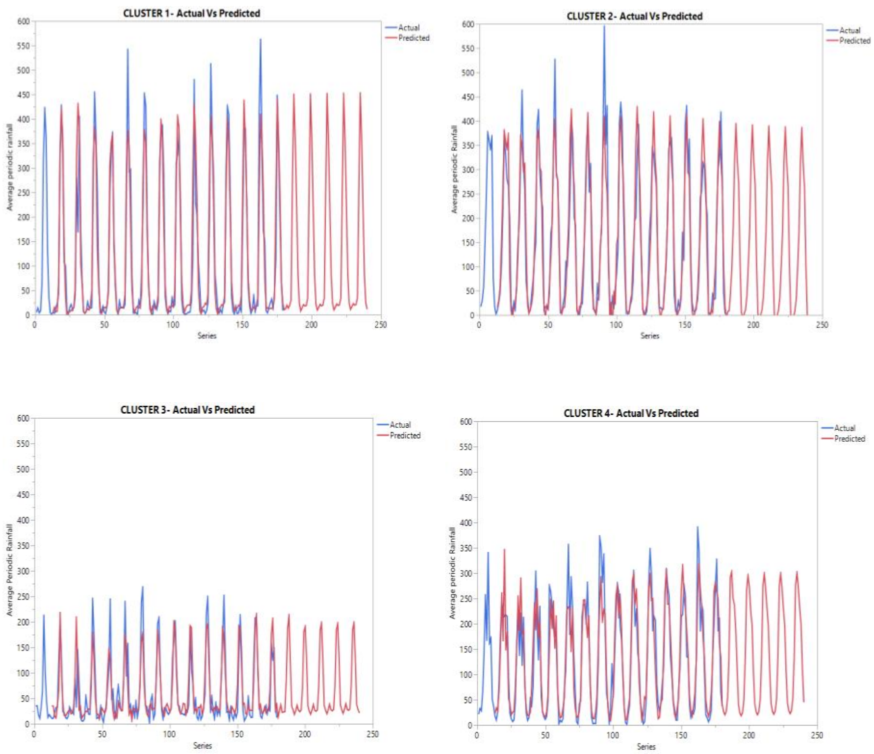



**Figure 18. Results Comparison: Actual Vs. Predicted**

We had obtained the actual annual rainfall data for 2015, which we had kept aside for our validation. Unfortunately, the 2016 data was not available to us. On comparing the actual and predicted 2015 results, we can see that the rainfall measures for cluster 2, 3 and 4 are quite accurate. Cluster 1 predictions are not as accurate as the other clusters. This is probably due to the historic uneven rainfall patterns observed in the northern parts of Karnataka, and Konkan areas of India.

| ANNUAL RAINFALL | CLUSTER 1 | CLUSTER 2 | CLUSTER 3 | CLUSTER 4 |
|---|---|---|---|---|
| ACTUAL 2015 | 1159.64 | 1880.56 | 815.03 | 1552.94 |
| PREDICTED 2015 | 1517.43 | 1761.49 | 819.72 | 1687.13 |
| PREDICTED 2016 | 1487.94 | 1730.35 | 790.33 | 1649.73 |
| PREDICTED 2017 | 1504.36 | 1712.91 | 810.96 | 1677.79 |
| PREDICTED 2018 | 1511.43 | 1695.81 | 818.25 | 1687.77 |
| PREDICTED 2019 | 1520.36 | 1678.80 | 829.11 | 1702.74 |

**Figure 19. Model Results**

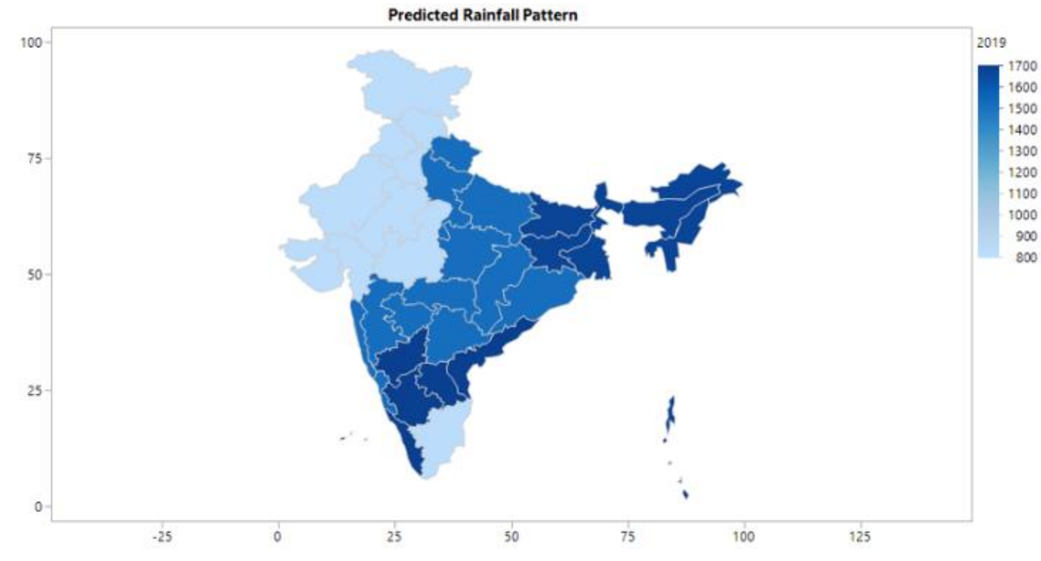



**Figure 20. Predicted Rainfall Pattern-2019**

The above map shows the predicted rainfall pattern and measures for India, for the year 2019.

## SUMMARY & INSIGHTS

The predicted rainfall for the next 2 years (2018 and 2019) is useful for policy makers and farmers. Some Kharif crops like Rice and Cotton require the right amount of rain water for maximum yield. If we consider the advantages of the forecasting, we can make full use of natural rainfall in the corresponding areas. These would provide a quantitative index and theoretical basis for the region to set a reasonable irrigation system. Apart from farming, monsoons in India influence the Indian Stock Markets. Barring other external factors, our findings can provide us with a rough estimate on how the stock markets will perform in the coming few months.

From our data analysis, we can see an upward trend in rainfall for Cluster Regions 1, 2 and 4 where as there is a dip in the annual rainfall for Cluster Region 3. It is highly possible that this dip is linked to the annual rise in temperature in the northern regions of India.

## CONCLUSION

With the help of Seasonal Arima modelling we were able to achieve forecasts of the average rainfall for the regions that were clustered based on their Time Series Similarity. This model can be further extended to forecast the average monthly rainfall of the individual subdivisions.

Seasonal ARIMA models can predict minimum and maximum temperature with good accuracy as statistics of models indicate. The accuracy of predictions made for rainfall by seasonal ARIMA model is less because data is abrupt, which increases white noise in the system. Another factor Seasonal ARIMA models cannot account for are natural weather occurrences like cyclones, tropical storms and extended heat waves.

## REFERENCES


1. Data Source: Open Government Data (OGD) Platform India (https://data.gov.in/catalog/all-india-area-weighted-monthly-seasonal-and-annual-rainfall-mm)
2. Rainfall Region Map of India (http://www.imd.gov.in/pages/rainfall_seasonal_cumulative_weekly.php)
3. Kharif Crops Analysis (https://agrinfobank.wordpress.com/2013/05/16/kharif-crops-list/)
4. JMP Custom Map Creator: Installer and Usage Instructions (https://community.jmp.com/t5/JMP-Add-Ins/Custom-Map-Creator/ta-p/21479)
5. Saeed Aghabozorgi,Ali Seyed Shirkhorshidi, TehYing Wah, Time-series clustering – A decade review, Volume 53, October–November 2015, Pages 16-38 (https://www.sciencedirect.com/science/article/pii/S0306437915000733)
6. Eamonn Keogh, Jessica Lin, Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research, Computer Science & Engineering Department University of California – Riverside
7. JMP Time Series Analysis: https://www.jmp.com/en_dk/events/ondemand/mastering-jmp/time-series-analysis-and-forecasting.html
8. Gary Fenga, Stacy Cobbb, Zaid Abdoc,f, Daniel K. Fisherd, Ying Ouyange, Ardeshir Adelia, and Johnie N. Jenkinsa: Trend Analysis and Forecast of Precipitation, Reference Evapotranspiration, and Rainfall Deficit in the Blackland Prairie of Eastern Mississippi (http://journals.ametsoc.org/doi/full/10.1175/JAMC-D-15-0265.1)
9. Pazvakawambwa G.T. and Ogunmokun A. A.: A time-series forecasting model for Windhoek Rainfall, Namibia, University of Namibia, Faculty of Engineering and IT (http://digitalcommons.andrews.edu/cgi/viewcontent.cgi?article=1146&context=arc)
10. Shengwei Wang, Juan Feng, Gang Liu: Application of seasonal time series model in the precipitation forecast, Mathematical and Computer Modelling, Volume 58, Issues 3–4, August 2013, Pages 677-683 (https://www.sciencedirect.com/science/article/pii/S089571771100639X)


## ACKNOWLEDGMENTS


The authors would like to express their sincere gratitude to Prof. Dr. Kam Tin Seong, Associate Professor of Information Systems (Practice)at the Singapore Management University (SMU) for his valuable support and advice. We also thank Open Government Data (OGD) Platform India for hosting the data.


## CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:


Name: Nevil Bruno
Enterprise: Singapore Management University
City: Singapore
Work Phone: +65 9373 3944
Mail: nevil.bruno.2017@mitb.smu.edu.sg

Name: Pankhuri Dwivedi
Enterprise: Singapore Management University
City: Singapore
Work Phone: +65 9422 1466
Mail: pankhurid.2017@mitb.smu.edu.sg

Name: Priyanka Sharma
Enterprise: Singapore Management University
City: Singapore
Work Phone: +65 8184 8966
Mail: priyankas.2017@mitb.smu.edu.sg